

# A Relevant Content Filtering Based Framework For Data Stream Summarization

Cailing Dong

Department of Information Systems  
University of Maryland, Baltimore County  
Baltimore, Maryland 21250  
Email: cailing.dong@umbc.edu

Arvind Agarwal

Palo Alto Research Center (PARC)  
Webster, New York 14580  
Email: arvind.agarwal@xerox.com

**Abstract**—Social media platforms are a rich source of information these days, however, of all the available information, only a small fraction is of users’ interest. To help users catch up with the latest topics of their interests from the large amount of information available in social media, we present a relevant content filtering based framework for data stream summarization. More specifically, given the topic or event of interest, this framework can dynamically discover and filter out relevant information from irrelevant information in the stream of text provided by social media platforms. It then further captures the most representative and up-to-date information to generate a sequential summary or event story line along with the evolution of the topic or event. Our framework does not depend on any labeled data, it instead uses the weak supervision provided by the user, which matches the real scenarios of users searching for information about an ongoing event. We experimented on two real events traced by a Twitter dataset from TREC 2011. The results verified the effectiveness of relevant content filtering and sequential summary generation of the proposed framework. It also shows its robustness of using the most easy-to-obtain weak supervision, i.e., trending topic or hashtag. Thus, this framework can be easily integrated into social media platforms such as Twitter to generate sequential summaries for the events of interest. We also make the manually generated gold-standard sequential summaries of the two test events publicly available<sup>1</sup> for future use in the community.

## I. INTRODUCTION

In the last few years, social media, in particular micro-blogging websites, has seen a steep rise in their popularity with the increasing number of users contributing the content in terms of short text messages. As of the 4th quarter 2015, there are about 305 million monthly active users on Twitter who posts more than 500 million Tweets everyday. When so much information coming at such a high speed, it is of vital importance to provide a concise and up-to-date summary to help users catching up and understanding the topic and event of interests. Besides the large quantity of the posts, these short text messages are status updates consists of various and ever-changing topics, ranging from simple status updates about personal life such as *going to visit a friend*, to text messages about ongoing real-life events such as *FIFA world cup*, *Justin Bieber performance*, to more involved conversations about the topics of general interests such as *global warming*, *terrorism*

etc. Although some of these posts are related to events of general interests, most are simply about daily-life routine events, and therefore, of little-to-no interest to a user other than the one who posted them, or to his immediate connections. In fact, a study shows that among all the posts on Twitter, only 3.6% of the posts are related to the topics of mainstream news [1]. Therefore, there is an apparent need for relevant content filtering, especially when the information is coming at such a high speed and with so much noise.

The usefulness of relevant content filtering system goes beyond the users searching for information on the website. This kind of system is useful for many end-user applications, especially if they were to operate on streaming data, such as sentiment analysis [2], summarization [3], topic detection[4], etc. Almost in all of the tasks that operate on Twitter data, the very first step should be removing uninteresting information. However, most of the work on micro-blogging summarization has not put as much effort on relevant content filtering before performing summarization. One major reason is the streaming nature of the data makes it hard for one to rely on a *static* method for information filtering. One instead needs a method that is able to filter content dynamically to track the evolving news and events.

In this paper, we present a relevant content filtering based framework for data stream summarization, namely **Weakly Supervised Stream Filter and Summarizer (WS<sup>2</sup>FS)**, which is suitable for both *relevant content filtering* and *sequential summarization*. Unlike classical supervised method that relies on the availability of labeled data, the proposed framework does *not* use any manually labeled data, it instead uses *weak supervision* from users, which can be as simple as topical keywords, or in the form of any rule that can provide global *feature-level* information. When using the most easy-to-obtain supervision, i.e., hashtags, our framework can be treated as an almost unsupervised method, making it practical to be used for summarizing both *personalized events* (i.e. events of personal interests) and *general events* (events of general interests). Another important strength of the framework is its ability to handle streaming data. The framework is modeled as an online classification framework, which evolves i.e., learns from the new data as it becomes available. Its independence on any labeled data and its ability to adapt to streaming data

<sup>1</sup>[https://drive.google.com/open?id=15jRw13i0xARUW3HqBn3BdR451Xk7P2Qj-HO\\_OFmMW0](https://drive.google.com/open?id=15jRw13i0xARUW3HqBn3BdR451Xk7P2Qj-HO_OFmMW0)

make it suitable for integration into streaming data websites such as Twitter, where it can generate up-to-date summaries for topics/events of interests. In general, our contribution in this paper is as follows:

- We propose a relevant content filtering based framework for data stream summarization. It couples the two important tasks in social media, i.e., *relevant content filtering* and *event summarization* in one integrated framework. This framework is not only able to capture the event evolution and dynamically filter out relevant content, but also generates sequential summary or event story line effectively.
- The proposed framework is almost unsupervised since it does not use any manual labeled data<sup>2</sup>. The *weak supervision* it uses can either be done automatically (making it fully unsupervised) or be provided by information seeker.
- It best simulates the real scenarios of users searching for relevant information with self-defined search queries from any social stream websites, hoping to get a concise and informative summary. Thus, it can be integrated into any social stream websites such as Twitter readily to generate the summaries of the event of interest in a timely fashion.
- Our experimental results showed that the hashtag-based *weak supervision* produces the best results. As such supervision is easy to obtain, our framework can be easily extended to generate both personal event summary and global event summary.
- We make the manually generated chunk-wise and sequential summaries of two real test events publicly available, which can be used readily in the community.

The remainder of the paper is structured as follows. Section II provides some related work. In Section III, we describe the general structure, major components and detailed implementation of the proposed framework. Section IV demonstrates the comprehensive experiments on two real events delivered in Twitter, and examines its performance using different types of weak supervisions. Finally we discuss and conclude the paper in Section V and Section VI.

## II. RELATED WORK

### A. Relevant Content Filtering

Most of the the work in content filtering has been based on the following two types of methods: (1) information retrieval based methods, (2) machine learning based method. In information retrieval based methods, a query is formed based on the information that is being sought, and then, the query is executed to find the relevant content. Although in theory, any traditional information retrieval based method can be used for this, streaming nature of the data on micro-blogging website makes it hard to implement. For online streaming social media content, using information retrieval method that employs pre-built indexes is not feasible. Although a mature field in itself, information retrieval field has not yet found its ground in

retrieving the content from streaming social media platforms. Most of the current work still relies on simple method such as query keywords based search. In machine learning based methods, a typical approach is to build topic specific *supervised* classifier [5]. However, these supervised classification methods have various limitations which makes them less appropriate for content filtering for streaming data. First of all, supervised classification methods need labeled data. Getting labeled data is both expensive and time consuming. Secondly, supervised classification methods are not easily extensible to new topics. Every time a new topic comes, one has to create new labeled data and then build a new classifier. Since the topics keep evolving in the data stream, it is not reliable to use a fixed labeled dataset to capture the whole event.

### B. Micro-blogging Summarization

Previous work on micro-blogging summarization can be divided into three categories, i.e., frequency-based methods [6], [7], graph-based methods [8], [9] and context-based methods [10], [11]. Frequency-based methods are based on the assumption that if a word or a set of words in a data instance (such as a tweet) has a high frequency of being repeated, the instances containing the set of high-frequency words must be good candidates for generating summary. Based on the similar assumption, graph-based methods build a word graph to capture common sequences of words about the given topic. The path with the highest total weight is regarded as a candidate summary instance. Typical graph-based methods include TextRank [12], LexRank [13] and Phrase Reinforcement (PR) algorithm [8]. Context-based approaches rate the importance of a data instance not only based on its textual importance, but also based upon other non-textual features, such as user influence, data instance popularity and temporal signals [10]. Although verified to be very effective in generating single-sentence summary, none of these algorithms were specifically designed for or have been used on streaming data. Furthermore, these methods are pure summarization methods assuming the relevant content is ready to use. Putting them into the streaming environment, the effectiveness of these summarization methods would not be guaranteed as they could fail to capture the evolution of the given topic based on *static* keywords.

In contrast, our proposed framework integrates both *relevant content filtering* and *summarization*. It is dynamic and changes its behavior according to the arriving data from the stream. We emphasize here that these summarization methods are not competitors to the proposed framework, they are rather complementary, i.e., any of these summarization methods can be integrated with the *relevant content filter* of our framework.

### C. Event Tracking and Summarization

Lately, event tracking and summarization has raised lots of attention, where one key task is to detect the relevant content about the event. One major application domain is summarizing scheduled events [14], [15]. For instance, Chakrabarti et al. [14] employed a modified Hidden Markov Model

<sup>2</sup>It is not explicitly labeled for the classification task, rather than obtained from the data itself.

(HMM) to learn the structure and vocabulary of multiple American football games, in order to detect relevant content with regard to future games and further summarize them. Nichols et al [15] used an unsupervised algorithms to generate summary for sporting events, in which relevant content were extracted by detecting spikes in volume of status updates and further ranking them using a phrase graph. Using the similar approach, Zubiaga achieved real-time summarization of scheduled sporting events [16]. Compared with scheduled events which usually have specific “moments” and terminology, tracking unscheduled events are more challenging, but of general applicability. In [17], Osborne et al. classified a tweet as relevant or not based on the score distribution within a set of tweets, and further generated summary by removing redundancy among the selected tweets. To build a large-scale corpus for evaluating event detection on Twitter, McMinn et al. [18] employed crowdsourcing to gather relevance judgments. Other work mostly focus on detecting data volume changes, extracting sub-topics and further clustering them into the same events [19], [20]. In contrast to these methods, our framework employs more sophisticated techniques grounded in machine learning (i.e. online learning) for filtering out relevant content from irrelevant one, and for summarizing events. Furthermore, our framework is designed for all events, scheduled or unscheduled.

### III. WEAKLY SUPERVISED STREAM FILTER AND SUMMARIZER (WS<sup>2</sup>FS)

To filter out the relevant content with regard to a given topic or event from a data stream, we need to classify each instance (e.g., each tweet) into “relevant” or “non-relevant” categories, which is a classic binary classification problem. A good classifier for streaming data needs:

- *Reliable training datasets.* For a text stream containing almost infinite set of topics, creating such training datasets through manual annotations is impractical. It calls for an automatic approach to creating reliable training datasets.
- *Maintain the “main thread” and capture the “evolution” of the event.* A good classifier for streaming data should be continuously learning. It should capture not only the main “theme” of the event, but also the content as the event unfolds, which is also an important feature of building a good summarizer on streaming data.

Motivated by the above two important tasks, we propose a **Weakly Supervised Stream Filter and Summarizer (WS<sup>2</sup>FS)** to filter out relevant content and further generate sequential summary from data stream. The general framework is shown in Figure 1.

#### A. Relevant Content Filter of WS<sup>2</sup>FS

The general structure of the *relevant content filter* in WS<sup>2</sup>FS is demonstrated at the top part of Figure 1. The first step is to define the appropriate size of a stream chunk and thereafter split the data stream into different data stream chunks. The size of the chunk represents the granularity of filtering. It is flexible to define, e.g., according to timestamp, data volume, etc. The

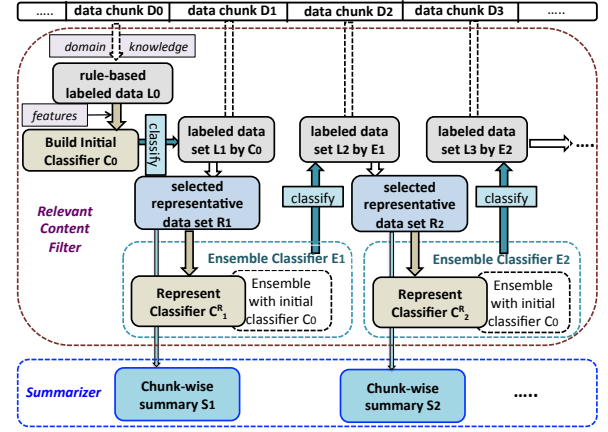


Fig. 1: Framework of WS<sup>2</sup>FS

next step is choosing an appropriate starting point. Although it can start from any time or any stream chunk, a good starting point is when the given event starts to emerge (similar to the spike in data volume), which is a good timing to capture the “main thread” of the event. In general, *relevant content filter* consists of three main components: (1) a one-time INITIAL CLASSIFIER  $C_0$  builder which builds  $C_0$  using the first stream chunk  $D_0$ , (2) a representative dataset  $R_i$  builder and (3) an ENSEMBLE CLASSIFIER  $E_i$  builder, for each chunk  $D_i$  ( $i \geq 1$ ). In the following, we describe each of these components in detail.

*First of all, build the initial classifier  $C_0$ .* Once the stream is split into chunks, an initial classifier  $C_0$  is built on first chunk  $D_0$ . Since  $D_0$  does not have labels, the very first task is to get labels for  $D_0$ . One major contribution of the proposed framework is that it does not use any manually labeled data. Instead, the labels of instances in  $D_0$  are created automatically based on *weak supervision* provided by the information seeker. We emphasize here that the labels are not obtained in a classical way, i.e., by asking label for each instance, rather, information seekers provide rules that operate on the whole corpora which in turn produce each label. This component results in a rule-based labeled dataset  $L_0$ , which is used to construct the INITIAL CLASSIFIER  $C_0$ .

*Second and third components correspond to building representative training datasets and classifiers for each of the following chunks.* These two components function alternatively, i.e., classifier in chunk  $i$  i.e.,  $C_i$  is used to classify the data in the next chunk i.e.,  $D_{i+1}$  (in other words get  $L_{i+1}$ ) which is then further processed to build the classifier for chunk  $i+1$ , i.e.,  $C_{i+1}$ . The training dataset for *each* chunk should be such that it captures the dynamic nature of the event in that chunk. In other words, it should contain any new content that appeared in that chunk. We build training dataset for chunk  $i+1$  using classifier from chunk  $i$ , followed by some further processing. This further processing consists of using *weak supervision* along with other available information in the dataset such as the number of followers in Twitter (detailed information will

be elaborated in the following part). We call such dataset as *chunk representative dataset* and denote it by  $R_i$ . The positive instances in  $R_i$  are selected in a way that they are highly reliable, and contain subtopics of the given event in chunk  $i$ . Thus they capture the “evolution” of the event and set up the up-to-date criteria for filtering relevant data from the next chunk.

In order to build a content filtering classifier for chunk  $i$ , we first build a classifier called REPRESENT CLASSIFIER  $C_i^R$  upon each  $R_i$ . As this classifier is built upon  $R_i$ , which mainly captures the localized subtopics that appeared in chunk  $i$ , it will fail to capture the main theme of the event. While, the main theme of the event is often captured by the first classifier  $C_0$ . In order to capture both the main theme of the event and the updated subtopics, for each chunk  $i$ , an ensemble classifier is built using two base classifiers, i.e., chunk-specific REPRESENT CLASSIFIER  $C_i^R$  and the INITIAL CLASSIFIER  $C_0$ . We call this classifier ENSEMBLE CLASSIFIER  $E_i$ . For each instance, its confidence of being a relevant instance is calculated by combining the confidence values produced by both of the two base classifiers, as given by:

$$Conf_{E_i}(x) = (1 - \alpha) * Conf_{C_0}(x) + \alpha * Conf_{C_i^R}(x). \quad (1)$$

The weight  $\alpha$  is flexible to set. If the provided *weak supervision* is not strong, we may need to put more weight on  $C_0$  to better maintain the main “theme”. On the contrary, if we are more interested in the evolution of the given event, we can put more weight on  $C_i^R$  to better capture the sub-topics.

*Chunk representative dataset  $R_i$  builder:* How to create *chunk representative dataset  $R_i$*  to build REPRESENT CLASSIFIER  $C_i^R$  is the key component of *relevant content filter* in WS<sup>2</sup>FS. As mentioned before, the positive instances  $R_i^+$  in  $R_i$  should capture the evolution of the given event. This evolution is usually captured by the words that most frequently co-occurred with the provided *weak supervision* or its induced rules. We call such words as *companion words*. It is worth noting that the purpose of the *weak supervision* goes beyond getting initial labels for  $D_0$ . They are used all along the text stream to obtain *companion words*.

The **qualification criteria  $\mathcal{Q}$**  for selecting representative dataset  $R_i$  is given by the pseudocode in Procedure 1. The selection of *companion words* (line 1-6) depends on the “3C” factors, i.e., *Confidence*, *Correlation* and *Credibility*. *Confidence* means the confidence of being relevant as judged by the ENSEMBLE CLASSIFIER  $E_i$ . *Correlation* measures the semantic relatedness to the given event. And *credibility* measures the reliability of the “source” of the instance. Specifically, line 1-2 make sure the candidate positive instances have high *confidence*. Line 3 and line 4 create two sorted list of instances according to *correlation* and *credibility* in descending order. The qualified instances are selected as the intersection of the top- $p$  instances in the two sorted lists, guaranteeing both high *correlation* and *credibility*. If the number of qualified instances is smaller than  $p$ , the top instances in the sorted list based on *correlation* are added as supplements (line 5). Finally, the top- $m$  most frequent words in the qualified instances are selected

---

### Procedure 1 Qualification Criteria $\mathcal{Q}$ of Selecting $R_i$

---

**Input:** Labeled data  $L_i$  from  $D_i$  and corresponding confidence value  $Conf_{E_i}$  ( $i \geq 0$ )

**Input:** #candidate instances =  $p$ ; #companion words =  $m$ ; #representative instances =  $n$

**Output:** *chunk representative dataset  $R_i$*  sorted by relevancy

▷ **Select companion words**

- 1:  $Avg^+(Conf_{E_i})$ : average *confidence* value of positive instances  $L_i^+$
- 2:  $L_i^{+'} \leftarrow \{x \in L_i^+ : Conf_{E_i}(x) \geq Avg^+(Conf_{E_i}(x))\}$
- 3:  $L_i^{C+} \leftarrow \text{sort } L_i^{+'} \text{ by } \textit{correlation} \text{ descendingly}$
- 4:  $L_i^{R+} \leftarrow \text{sort } L_i^{+'} \text{ by } \textit{credibility} \text{ descendingly}$
- 5:  $L_i^{Cand+} \leftarrow (\text{top-}p(L_i^{C+}) \cap \text{top-}p(L_i^{R+})) \cup [\text{top}(L_i^{C+})]$
- 6:  $Comp_i \leftarrow \text{top-}m \text{ most frequent words in } L_i^{Cand+}$

▷ **Select distant words**

- 7:  $Avg^-(Conf_{E_i})$ : average *confidence* value of negative instances  $L_i^-$
- 8:  $L_i^{-'} \leftarrow \{x \in L_i^- : (Conf_{E_i}(x) \leq Avg^-(Conf_{E_i}(x)))\}$
- 9:  $L_i^{C-} \leftarrow \text{sort } L_i^{-'} \text{ by } \textit{correlation} \text{ ascendingly}$
- 10:  $L_i^{R-} \leftarrow \text{sort } L_i^{-'} \text{ by } \textit{credibility} \text{ ascendingly}$
- 11:  $L_i^{Cand-} \leftarrow (\text{top-}p(L_i^{C-}) \cap \text{top-}p(L_i^{R-})) \cup [\text{top}(L_i^{C-})]$
- 12:  $Distant_i \leftarrow \text{top-}m \text{ most frequent words in } L_i^{Cand-}$

▷ **Select representative instances and build  $R_i$**

▷ **check diversity**

- 13:  $Common_i = Comp_i \cap Distant_i$
- 14: **if** ( $\frac{|Common_i|}{|Comp_i|} < 0.5$ ) & ( $\frac{|Common_i|}{|Distant_i|} < 0.5$ ) **then**
- 15:  $Comp_i \leftarrow Comp_i \cup T$
- 16:  $S_i^+ \leftarrow \text{top-}n(L_i^{+'}) \text{ on frequency of } Comp_i$
- 17:  $S_i^- \leftarrow \text{top-}n(L_i^{-'}) \text{ on frequency of } Distant_i$
- 18:  $R_i \leftarrow S_i^+ \cup S_i^-$
- 19: **end if**

---

as *companion words* (line 6). In order to build REPRESENT CLASSIFIER, we also need to select representative negative instances. The procedure is described in line 7-12 in Procedure 1, which is based on low values of *confidence*, *correlation*, and *credibility*. We call the words selected from these negative instances as *distant words*. Finally, the instances that have highest frequency of *companion words* and *distant words* are selected as representative positive and negative instances, respectively (line 16-17). Before doing this, we need to check the *diversity* of the two sets of words (line 14). If the ratio of the size of the common words to the size of either set is low, it means the two sets of words have large diversity and therefore are reliable enough to be selected as representative instances. Otherwise, the representative dataset building process will be skipped from this given stream chunk.

### B. Summarizer of WS<sup>2</sup>FS

The *summarizer* of WS<sup>2</sup>FS is seen at the bottom of Figure 1. For each chunk, the *relevant content filter* in WS<sup>2</sup>FS has produced a list of representative positive instances  $R_i^+$ , sorted based on *confidence*, *correlation* and *credibility*. The final ranking signifies the importance and representativeness of the instances. Thus, the top-ranked instances in each stream chunk can be regarded as good candidates to generate the *chunk-wise summary* of the given event. Later, the *summarizer* of WS<sup>2</sup>FS combines all the chunk-wise summaries in chronological order, and further process it to generate the final *sequential summary*.

A key point here is how to select a candidate data instance as a final chunk-wise or sequential summary instance. As the

resultant summary should cover as many aspects of the event as possible, we use *diversity* as the selection measurement. *Diversity* refers to the opposite of redundancy here. That is, all the instances should minimally overlap with each other within the final *chunk-wise summary* and *sequential summary*. In our work, we employ ROUGE-L [21] to calculate the degree of overlapping between a candidate instance and each of the already chosen summary instances. ROUGE-L calculates the statistics (average recall, precision and F1 values) about the longest common subsequence between two string. Obviously, the higher the value is, the less diversity the candidate instance will bring. A threshold  $\theta_{diversity}$  can be set flexibly based on the desired level of diversity. In our work, to keep higher diversity, we regard a candidate instance to be a final summary instance when all the three statistic values are less than 0.4.

#### IV. EXPERIMENTS

Twitter is a representative source of data stream. In our experiment, we apply WS<sup>2</sup>FS on a Tweet stream dataset TWEETS2011 from TREC 2011 Microblog Track<sup>3</sup>. The dataset contains 16 million tweets sampled between January 23rd and February 8th, 2011. It also provides a set of manually labeled relevant tweets for 50 topics. Two important events happened during the two weeks of sampled tweets, *Moscow airport bombing* and *Egyptian revolution*. In this dataset, the first event is described by one topic *Moscow airport bombing*, while the second event covers three topics, *Egyptian curfew*, *Egyptian evacuation* and *Egyptian protesters attack museum*. We combine all the tweets related to the three topics and associate them with event *Egyptian revolution*. As the original annotation is based on both the tweet text and the content of the URL [22], we asked two annotators to re-annotate the test dataset only based on tweet text. The inter-annotator agreement is 0.952 measured by Cohen’s kappa coefficient.

##### A. Experimental Settings

(1) *Stream chunks*. We split the data stream into different chunks by the creation “date” of the tweets.

(2) *Weak supervision*. In most cases, the initial knowledge of the users about an event is a general concept. Therefore, we simply define the *weak supervision* as: “A tweet is a relevant instance if any topical words (case insensitive) appear in the tweet”. We also treat the results of applying the *weak supervision* on the testing dataset as our *Baseline* method, which is also a simulation of relevant content filtering via Twitter Search API.

In practice, different users may provide different types/levels of *weak supervision* about the event. As event *Egyptian revolution* covers three aspects in the original datasets, we simulate the scenarios of users having the following three types of *weak supervision* about this event:

- *Type-a: general concept* – when a user cares about the event in general, described by topical words “Egyptian, revolution”;

- *Type-b: trending topic/hashtag* – when a user is interested in the trending topic or hashtag “#Jan25”;
- *Type-c: specific aspects* – when a user is interested in specific aspects of the event, described by the keywords “Egyptian, protesters, attack, museum, curfew, evacuation” (these keywords are chosen based on the topical words in the three topics in the original dataset).

(4) *Features*. WS<sup>2</sup>FS is designed as a general framework that does not rely on any specific features, thus we only use the following three type of features:

- *Keywords-freq*: the total frequency of topical key words in the tweet text. It is used to measure one of the “3C” factors, i.e., *correlation*.
- *Status*: We define status as the normalized ratio between the number of followers and followings of the user who wrote or retweeted the current tweet. The *status* shows the reputation of the tweet author to some extent. Intuitively, the higher the value of *status*, the more reliable is the tweet. We use it to measure the *credibility* in qualification criteria  $\mathcal{Q}$ .
- *Content-words*: the words in tweet text that carry the content. We use Twitter NLP toolkit<sup>4</sup> [23] to get part-of-speech (POS) tag for each word in tweet text. Only the content words with specific POS tags (“N”, “”, “S”, “Z”, “V”, “A”, “R”, “D”) are kept.

(5) *Starting point*. As described in Section III-A, a good and meaningful starting point in the stream is when the given event starts to emerge. Thus, we choose Jan 23rd, 2011 and Jan 25th, 2011 as the starting point to collect the first stream chunk for events *Moscow airport bombing* and *Egyptian revolution*, respectively.

(6) *Classifier*. As Naïve Bayes has been verified to be very effective in many text mining related tasks, we use it to build our classifiers. In addition, we set  $\alpha$  in Equation 1 to be 0.5 and 0.8 for events *Moscow airport bombing* and *Egyptian revolution* respectively, as the later involves many sub-events we want to capture.

(7) *Other parameters*. As the number of topical words-related instances in each chunk is less than 2000 on average, we choose to set small numbers on the parameters in Procedure 1. By teasing on different values on these parameters, we finally settled on selecting around 10% of the instances as candidate instances, among which only 10% of the instances are chosen as representative instances, in order to avoid topic drifting.

##### B. Performance of Relevant Content Filter

To demonstrate the performance of *relevant content filter* of WS<sup>2</sup>FS, we focus on its fundamentals, i.e., *companion words*, and its performance on the fixed testing dataset. In Table I, we showed the companion words in event *Moscow airport bombing* where the words in bold are event-relevant ones. We can find that at the early stage of the framework, the companion words are closely related to this event. As the topic/event starts to die out or submerge by new topics,

<sup>3</sup><http://trec.nist.gov/data/tweets/>

<sup>4</sup><http://www.ark.cs.cmu.edu/TweetNLP/>

TABLE I: Companion words in event *Moscow airport bombing*

| ChunkID | Companion words  |
|---------|--|
| Jan-24  | injured, killed, blast, explosion, 31, dead, bombing, suicide          |
| Jan-25  | blast, news, 35, killed, terrorist, attack, bombing, suicide, dead     |
| Jan-26  | modern, revenge, russian, heathrow, girlfriend, video, bombing         |
| Jan-27  | san, russian, news, bombing, call, police, ap, sings                   |
| Jan-28  | international, san, gatwick, blvd, domodedovo, terror, attack, bombing |
| Jan-29  | malaga, blast, shut, investigators, orlando, latest, bombing           |
| Jan-30  | san, international, passengers, stream, watch, security, people, live  |
| Jan-31  | san, london, news, subject, bomber, introduce, luton, #egypt           |
| Feb-02  | source, russian, victim, dfw, international                            |
| Feb-03  | cairns, open, townsville, richmond, international, opening, security   |
| Feb-07  | islamist, san, umarov, ordered, guardian, doku, operators, rebel       |
| Feb-08  | claims, umarov, ordered, leader, doku, rebel, bombing, chechen         |

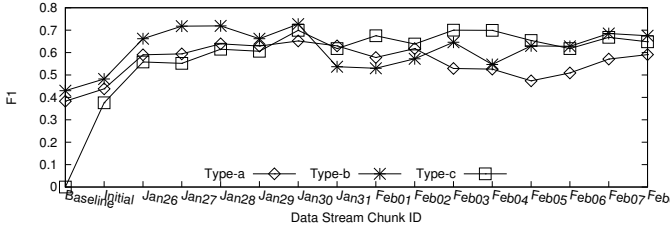


Fig. 2: Performance of *relevant content filter* on event *Egyptian revolution* with different types of *weak supervision*.

the selected companion words starts to slightly drift from the main theme. However, when there are updated information, the given topic/event comes alive again. Overall, the companion words selected by *relevant content filter* can capture the evolution of the events to a large extent.

In Figure 2, we show the F1 score of *content relevant filter* on the testing dataset of event *Egyptian revolution*, with three types of *weak supervision* defined earlier. As we can see, based on *Type-a weak supervision*, it retrieves the relevant content quite well, mainly because general concept is universally acknowledged and widely used to discuss about this event. However, the performance begins to decrease as the event evolves, which is probably because people tend to discuss more detailed aspects of the event along its evolution. Armed with the hashtag based *Type-b weak supervision*, the *relevant content filter* filters out the relevant content with high accuracy in the first few chunks. Later on, as the event evolves, more dynamic “labels” are created to describe the event, so the performance simply based on the hashtag slightly decreases in some chunks. Overall, the trending topic/ hashtag born with the event can capture the main theme of the event along with its lifetime. As shown in this figure, the F1 score generated by *Baseline method* with *Type-c weak supervision* is around 0. This is because it simply classifies all the tweets containing any of those keywords as relevant content. But later on, our *content relevant filter* produces much more accurate relevant content as the event evolves. In general, the *relevant content filter* of *WS<sup>2</sup>FS* can effectively capture the dynamically changing relevant content of an event, with different levels of *weak supervision*.

TABLE II: Quality of *chunk-wise summary* on event *Moscow airport bombing* (average value).

|           | Centroid | LexRank | Q1    | Q2    | Q3    | WS <sup>2</sup> FS |
|-----------|----------|---------|-------|-------|-------|--------------------|
| Precision | 0.176    | 0.247   | 0.518 | 0.607 | 0.586 | <b>0.706</b>       |
| Recall    | 0.235    | 0.167   | 0.385 | 0.374 | 0.316 | <b>0.726</b>       |
| F1        | 0.200    | 0.195   | 0.434 | 0.449 | 0.394 | <b>0.714</b>       |

### C. Performance of Summarizer

*Comparison summarization methods:* To compare the *summarizer* of *WS<sup>2</sup>FS*, we choose the following commonly used classical summarization methods as baseline methods:

- (1) *Centroid*: the centroid instance in the dataset is chosen to be the candidate summary instance [24].
- (2) *LexRank*: each instance is modeled as a vertex and the edges are created based on the cosine similarity of the TF-IDF vectors of the two vertices. Graph ranking method (e.g. PageRank) is used to select candidate summary instances [13].
- (3) *Query-based method*: from the perspective of information retrieval, the summary instance is chose from the most relevant documents to the given query. It contains the following three different kinds of similarity measures:

- *QueryCosine (Q1)*: uses cosine similarity of the TF-IDF vectors.
- *QueryCosineNoIDF (Q2)*: uses cosine similarity on TF vectors.
- *QueryWordOverlap (Q3)*: uses the overlapping of uni-grams in both document and query to measure the similarity.

*Evaluation criteria for generated summaries:* We compare these baseline methods with the *summarizer* of *WS<sup>2</sup>FS* on both *chunk-wise summary* and *sequential summary*. To guarantee the quality of the gold-standard summaries, we ask the annotators to grasp the “big moments” along the timeline of the events. Specifically, we extract some facts happened on different dates about the event *Moscow airport bombing*, based on its Wikipedia page<sup>5</sup>. For event *Egyptian revolution*, the annotations should capture the key facts follow the timelines provided by both its Wikipedia page<sup>6</sup> and the report from Al Jazeera English<sup>7</sup>. The two annotators manually choose the top-3 (if available) most important tweets in each stream chunk as the gold-standard *chunk-wise summary*. Accordingly, using the selection criterion *diversity* described in Section III-B, the top-3 tweets with highest score generated by *summarizer* of *WS<sup>2</sup>FS* and all the baseline methods are regarded as their *chunk-wise summaries*. Each of the two annotators also manually creates a *sequential summary* along with all the chunks. Using the same selection criterion *diversity*, the *summarizer* of *WS<sup>2</sup>FS* and each of the baseline methods also generate their own *sequential summaries* for both events.

*Quality of chunk-wise summary:* In Table II, we show the quality of *chunk-wise summary* generated by different summarization methods on event *Moscow airport bombing*.

<sup>5</sup>[http://en.wikipedia.org/wiki/Domodedovo\\_International\\_Airport\\_bombing](http://en.wikipedia.org/wiki/Domodedovo_International_Airport_bombing)

<sup>6</sup>[http://en.wikipedia.org/wiki/Timeline\\_of\\_the\\_Egyptian\\_Revolution\\_of\\_2011](http://en.wikipedia.org/wiki/Timeline_of_the_Egyptian_Revolution_of_2011)

<sup>7</sup><http://www.aljazeera.com/news/middleeast/2011/01/201112515334871490.html>

TABLE III: Quality of *chunk-wise summary* on event *Egyptian revolution* (average value).

|           | Centroid | LexRank  | Q1       | Q2       | Q3       | WS <sup>2</sup> FS |
|-----------|----------|----------|----------|----------|----------|--------------------|
| Precision | 0.272(b) | 0.232(c) | 0.192(c) | 0.203(c) | 0.205(c) | <b>0.597(b)</b>    |
| Recall    | 0.277(b) | 0.287(c) | 0.359(b) | 0.274(c) | 0.336(b) | <b>0.576(b)</b>    |
| F1        | 0.277(b) | 0.252(c) | 0.242(c) | 0.229(c) | 0.245(c) | <b>0.585(b)</b>    |

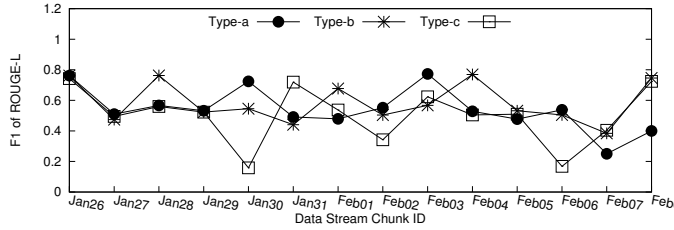


Fig. 3: Quality of *chunk-wise summary* on event *Egyptian revolution* in each chunk, produced by *summarizer* of WS<sup>2</sup>FS with different types of *weak supervision*.

Due to the space limitations, we only list the average values across all chunks. As we can see, our *summarizer* of WS<sup>2</sup>FS produces the best results in terms of *precision*, *recall* and *F1*. That is, the chunk-wise summary generated by WS<sup>2</sup>FS are most similar to the manually generated summary.

In Table III, we demonstrate the best average ROUGE-L scores of chunk-wise summarization produced by corresponding *weak supervision* type (indicated in the parenthesis) on event *Egyptian revolution*. On the whole, our *summarizer* produces the best results using *Type-b* supervision. For the baseline methods, the best results are usually generated using *Type-b* and *Type-c* *weak supervision*. This is because they largely rely on topical words, and can not fully explore the general concept (*Type-a*).

Furthermore, we are interested in investigating how different types of *weak supervision* affect the performance of our *summarizer* in producing summary in each chunk along with the event evolution. The F-1 values on each chunk are shown in Figure 3. In general, the performance fluctuates more with *Type-c* *weak supervision*, probably due to the matching “degree” between its keywords and the corresponding big moments of the event. In the real scenario, what keywords will be associated to an unscheduled event is nearly unable to predict. Thus, the method which can generate good summary based on hashtag (*Type-b*) or general information (*Type-a*) is more useful than the method which rely on specific topical words (*Type-c*), which again verified the practical usefulness of our proposed framework which makes better use of *Type-b* supervision.

*Results of sequential summarization:* Table IV shows the quality of sequential summaries generated by different methods on event *Moscow airport bombing*, compared with the the gold-standard sequential summary. From this table, we can see that *query-based* methods tend to produce higher precision but lower recall. *Centroid* and *LexRank* produce poor results overall. Our *summarizer* produces the best results. The evaluation of sequential summarization on event *Egyptian*

TABLE IV: Quality of *sequential summary* on event *Moscow airport bombing*.

|           | Centroid | LexRank | Q1    | Q2           | Q3    | WS <sup>2</sup> FS |
|-----------|----------|---------|-------|--------------|-------|--------------------|
| Precision | 0.267    | 0.439   | 0.544 | <b>0.682</b> | 0.649 | 0.631              |
| Recall    | 0.338    | 0.177   | 0.416 | 0.457        | 0.374 | <b>0.662</b>       |
| F1        | 0.299    | 0.252   | 0.471 | 0.547        | 0.474 | <b>0.646</b>       |

*revolution* are shows in Figure 4, with different methods under different types of *weak supervision*. It shows the similar results as of the chunk-wise summaries in Table III. That is, WS<sup>2</sup>FS produces the best results. In terms of *weak supervision*, those baseline methods benefit more from *Type-c* *weak supervision* as they rely more on topical words. WS<sup>2</sup>FS produces the best results with hashtag-based *Type-b* *weak supervision*, and generates good enough results based on *weak supervision* coming from general concept (*Type-a*) as well. The resultant sequential summary produced by the *summarizer* on the two events are publicly available along with the gold standard summaries<sup>1</sup>.

## V. DISCUSSION

One of the important inputs to our system is *weak supervision*, this weak supervision should not only be easy to obtain for a variety of events but also be effective in generating meaningful summaries. In our experiments, we have shown that it is possible to achieve both goals simultaneously. More specifically, we have experimented and shown results for both *relevant content filtering* and *summarization* based on three different types of *weak supervision* that are relatively easier to obtain. Among these three types of weak supervision, *type-b* weak supervision (i.e. trending-topic/hashtag based) has performed the best. This hashtag-based weak supervision also happen to be the one that can be most easily obtained. Other than its easy availability, hashtag-based weak supervision also has an immediate practical advantage. For the users who tweet with a particular hashtag, the proposed summarization method can be used to provide them an up-to-date summary of the events related to their hashtags, which is a very useful feature to be integrated in social media. In addition to the personalized event summary, the framework can also be used to provide global event summary based on the hashtags trending on the site. Such advantage demonstrates the practicality of our method and strengthens the argument for its adaptation into real world applications. While our method is able to exploit such power of trending topics, unfortunately those baseline methods cannot do so. If one were to treat the initial weak supervision as a query, then our method is able to produce the dynamic query-based summary (i.e. evolution of the event) as opposed to the static summaries produced by the baselines. It is important to note that our method relies on trending topics only for initial weak supervision, and updates itself as the new data comes in, unlike other query based summarization methods which entirely depend on the initial query to produce summaries.



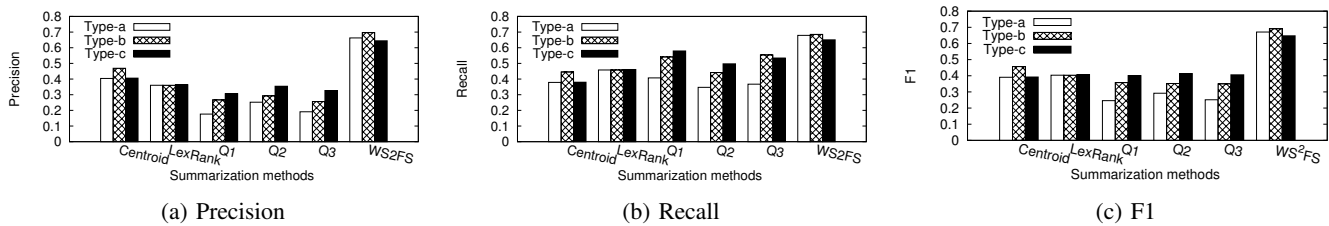


Fig. 4: Quality of *sequential summary* on event *Egyptian revolution* with different types of *weak supervision*.

## VI. CONCLUSION

In this paper, we have presented a relevant content filtering based framework for data stream summarization in social media platforms. This framework does not use any labeled data as typical supervised methods do; it instead uses the weak supervision in the forms of rules and guidelines. Such weak supervision makes this framework extensible for different applications. In addition, our framework is built for streaming environment, i.e. it is not only able to filter out the relevant content with regard to the given topic or event, but also can generate the sequential summary or event story line as it evolves in the text stream. Our experiments on a set of tweets about two real events verified its effectiveness in generating summarization from data stream after relevant content filtering. Besides, our experimental results showed that using the most easy-to-obtain weak supervision, i.e., hashtags, it can generate the best results. Such property makes our framework more applicable: it can be regarded as an almost unsupervised framework given the trending topic or hashtag; it can be easily used to generate both personal event summary and global event summary, based on the personal defined hashtags and trending hashtags representing global events, respectively. Thus, our framework can be easily integrated into social media platforms such as Twitter.

## REFERENCES

- [1] R. Kelly, "Twitter study – august 2009," August 2009. [Online]. Available: <http://pearanalytics.com/wp-content/uploads/2012/12/Twitter-Study-August-2009.pdf>
- [2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, 2011, pp. 30–38.
- [3] M. Krieger and D. Ahn, "Tweetmotif: Exploratory search and topic summarization for twitter," in *ICWSM*, 2010.
- [4] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012, pp. 536–544.
- [5] M. A. H. Khan, M. Iwai, and K. Sezaki, "An improved classification strategy for filtering relevant tweets using bag-of-word classifiers," *Journal of Information Processing*, vol. 21, no. 3, pp. 507–516, 2013.
- [6] B. Sharifi, M.-A. Hutton, and J. K. Kalita, "Experiments in microblog summarization," in *2010 IEEE Second International Conference on Social Computing*, 2010, pp. 49–56.
- [7] A. Olariu, "Hierarchical clustering in improving microblog stream summarization," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2013, pp. 424–435.
- [8] B. Sharifi, M.-A. Hutton, and J. Kalita, "Summarizing microblogs automatically," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 685–688.
- [9] A. Olariu, "Efficient online summarization of microblogging streams," *EACL 2014*, p. 236, 2014.
- [10] Y. Chang, X. Wang, Q. Mei, and Y. Liu, "Towards twitter context summarization with user influence models," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 527–536.
- [11] X. Yang, A. Ghoting, Y. Ruan, and S. Parthasarathy, "A framework for summarizing and analyzing twitter feeds," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 370–378.
- [12] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.
- [13] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.(JAIR)*, vol. 22, no. 1, pp. 457–479, 2004.
- [14] D. Chakrabarti and K. Punera, "Event summarization using tweets." in *ICWSM*, 2011.
- [15] J. Nichols, J. Mahmud, and C. Drews, "Summarizing sporting events using twitter," in *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*. ACM, 2012, pp. 189–198.
- [16] A. Zubiaga, D. Spina, E. Amigó, and J. Gonzalo, "Towards real-time summarization of scheduled events from twitter streams," in *Proceedings of the 23rd ACM conference on Hypertext and social media*. ACM, 2012, pp. 319–320.
- [17] M. Osborne, S. Moran, R. McCreddie, A. Von Lunen, M. Sykora, E. Cano, N. Ireson, C. Macdonald, I. Ounis, Y. He *et al.*, "Real-time detection, tracking, and monitoring of automatically discovered events in social media," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 37–42.
- [18] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013, pp. 409–418.
- [19] R. Long, H. Wang, Y. Chen, O. Jin, and Y. Yu, "Towards effective event detection, tracking and summarization on microblog data," in *Web-Age Information Management*. Springer, 2011, pp. 652–663.
- [20] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller, "Twitinfo: aggregating and visualizing microblogs for event exploration," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2011, pp. 227–236.
- [21] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.
- [22] J. L. Iadh Ounis, Craig Macdonald and I. Soboroff, "Overview of the trec-2011 microblog track," in *Proceedings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.
- [23] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 42–47.
- [24] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.